From Show to Tell: A Survey on Deep Learning-Based Image Captioning

Year: 2021 | Citations: 339 | Authors: Matteo Stefanini, Marcella Cornia, L. Baraldi, Silvia Cascianelli, G. Fiameni

---

Abstract

Connecting Vision and Language plays an essential role in Generative Intelligence. For this reason, large research efforts have been devoted to image captioning, i.e. describing images with syntactically and semantically meaningful sentences. Starting from 2015 the task has generally been addressed with pipelines composed of a visual encoder and a language model for text generation. During these years, both components have evolved considerably through the exploitation of object regions, attributes, the introduction of multi-modal connections, fully-attentive approaches, and BERT-like early-fusion strategies. However, regardless of the impressive results, research in image captioning has not reached a conclusive answer yet. This work aims at providing a comprehensive overview of image captioning approaches, from visual encoding and text generation to training strategies, datasets, and evaluation metrics. In this respect, we quantitatively compare many relevant state-of-the-art approaches to identify the most impactful technical innovations in architectures and training strategies. Moreover, many variants of the problem and its open challenges are discussed. The final goal of this work is to serve as a tool for understanding the existing literature and highlighting the future directions for a research area where Computer Vision and Natural Language Processing can find an optimal synergy.